DOCUMENT RESUME

ED 177 221 TM 009 777

AUTHOR Bligh, Harold F.
TITLE Achievement Testing--A Look at Trends.
PUB DATE 10 Apr 79
NOTE 16p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, California, April 9-11, 1979)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; Cognitive Measurement; Curriculum Development; Educational Improvement; *Educational Testing; *Educational Trends; Elementary Secondary Education; Information Dissemination; Norms; Scores; *Standardized Tests; Test Bias; Test Construction; *Testing Problems; Test Interpretation

ABSTRACT

The strengths and weakness of standardized tests, and trends in achievement testing in the last 15 years are examined. The discussion of achievement tests includes survey, instructional, diagnostic, and basic skills tests, as well as tests used for formative and summative evaluation. Minimum competency tests are not examined in detail. Advantages of the tests are summarized: accuracy, objectivity, comparability, convenience, and efficiency. Multilevel achievement test series recognized as providing comprehensive and continuous measurement from kindergarten through high school. Critics are concerned with procedures for administering, interpreting, and using tests; the emphasis on measurement of cognitive skills; the emphasis on recall--rather than analysis, synthesis, and other higher order skills; the reliance on multiple-choice and/or poorly constructed test items; the failure to measure what is being taught; and test bias. It is asserted that criticism has contributed to the improvement of achievement tests. In test development, attention is being paid to: eliminating bias from test items; providing self-scoring or latent image printing tests for more immediate feedback; involving teachers and educational personnel in writing objectives and test items; developing item banks; providing more interpretive and prescriptive information; developing multilevel, scaled score systems; using latent trait models; and providing expectancy tables. The impact of the impetus for truth in testing is still unknown. (MH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED177221

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ACHIEVEMENT TESTING - A LOOK AT TRENDS

Harold F. Bligh
The Psychological Corporation

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Harold F. Bligh

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Presented at:

1979 Convention
National Council on Measurement in Education

San Francisco, California - April 10, 1979

TM 009 777

ACHIEVEMENT TESTING - A LOOK AT TRENDS

Harold F. Bligh
The Psychological Corporation

The past five to ten years have been marked by concerns that achievement tests no longer meet the needs for which they were originally developed. Teacher-made tests, contracts with students and other approaches have been offered as alternatives (Quinto and McKenna, 1977). Criterion-referenced testing has been seen by some as the panacea for the shortcomings of norm-referenced tests. In spite of the turmoil, the demands for standardized achievement measures continue, some new and others relatively unchanged or cast in different forms. These demands are being listened to by test authors and publishers and responses to them are already reflected in some of the more recently published instruments.

In this paper, I intend to look at some of the advantages of standardized achievement tests and at concerns about their shortcomings. I will then attempt to relate these to trends in achievement testing over the past decade and a half. First I would like to make three points that I think are relevant to today's discussions. When one speaks of a standardized achievement test, the concept of the traditional survey instrument sampling a broad domain of content and built to yield a normal distribution of scores for a prescribed grade or other instructional group may immediately come to mind. However, standardized achievement tests of today must be viewed in a much broader context -- that is, as including survey, instructional and diagnostic and a variety of basic skills tests, tests that are used for formative as well as for summative evaluations. Several of the more recently published diagnostic and instructional tests provide normative as well as objective-referenced or criterion-referenced information. These I am including within the broader context; but, I am not considering minimum competency tests, in particular those used in setting standards for high school graduation. Persons

interested in information on this latter topic will find excellent sources in the articles by Glass (1978) and others in the JEM 1978 Winter issue, and the Haney and Madaus report, "Making Sense of the Competency Testing Movement" (1978). The bibliographies accompanying these papers are extensive and include reports and other treatises as well as recently published articles.

In assessing the strengths and weaknesses of standardized achievement tests, we need also to keep in mind that their primary purpose is to provide relevant information to be used with other sources in decision making. All too often, the critics assume that achievement test scores are being used as the sole criterion for decision making in the schools and seem unaware of the concerted efforts of authors, publishers and other measurement specialists to alert test users to the fact that test scores are but one source for making judgments about pupil or group performances.

The third point is that strengths and weaknesses of achievement tests do not appear as distinct entities; usually they are interrelated and impact on the development as well as on the interpretation and use of tests. The "bias" issue, for example, impacts on the kinds of information that the test is to provide, the selection of content to be included, the types of pictures and item types used in eliciting responses, editing and field testing of the items, selection of normative populations, and the way the tests are administered, and the uses made of the results.

Now, what are some of the advantages of standardized achievement tests? In an address "Test Scores in Perspective" delivered at the 1978 National Conference on Achievement Testing and Basic Skills sponsored by the Department of Health, Education and Welfare, Turnbull (1978) listed accuracy, objectivity and comparability as the three basic characteristics contributing to the early and continued acceptance of standardized achievement tests. To these I would add convenience and efficiency.

In addition to the generally recognized advantages of standardized tests, multilevel achievement test series provide for comprehensive and continuous measurement from kindergarten through high school. When multilevel batteries of tests in the major school subjects are developed simultaneously, they provide vertical coordination through well defined content strands and multilevel scaling. Content appropriate for assessing performance at one level is related to that which comes before and after in the instructional sequence. Comparable normative groups provide for meaningful interpretation from level to level. Both of these features enhance the flexibility of the tests for on grade and out-of-level testing as well as for other nonstandard uses. Coordination within battery level provides for profile interpretations across subject matter areas. The availability of comparable and equivalent forms developed at the same time and according to the same content and statistical specifications is a major advantage.

Although standardized achievement tests are seen by many as serving their purposes well, attacks against them continue to appear in educational journals, as well as in newspapers and magazines. Concerns that lead to the National Educational Association's recommendation encouraging the elimination of group standardized intelligence, aptitude and achievement tests are presented by Quinto and McKenna (1977) in the NEA Handbook "Alternatives to Standardized Testing". The NEA concerns are representative of those that have been presented by others. Although concerned with procedures for administering, interpreting and using tests, these educators see even greater problems with the content and structure of the tests. They are concerned that standardized tests measure cognitive learnings almost totally; tap recall to the neglect of analysis, synthesis and other higher order skills and abilities; rely heavily on the use of multiple-choice and irrelevent and poorly constructed items; fail to measure much of what is being taught in the classroom; and finally are culturally biased.

In his perspective on testing, Turnbull (1978) identified three fallacies that may have encouraged the misuse of achievement test scores. These he categorized as the micrometer fallacy, that is, the investing of test scores with a precision that they never were meant to possess; the whole person fallacy, or the tendency to regard achievement tests as going beyond the measurement of academic accomplishments to tap attitudes and other characteristics of the "whole" person; and the equal preparation fallacy, the expectation that tests should somehow compensate for the differences in backgrounds that children bring to the testing situation. Many of the concerns expressed by critics of standardized tests fall into one or more of those three categories.

During the past several years, articles taking a more positive view of testing have been appearing, some as rebuttals to the earlier criticisms and others as attempts to clarify the issues and to bring about better understanding of the reasons for giving tests, of the ways tests are built and of the interpretations and uses that can appropriately be made of them. NCME is making a real contribution through the Measurement in Education reports. Of particular relevance to today's discussions are "The Paradox of Educational Testing," "On Reporting Test Results to Community Groups," "Using Your Achievement Test Score Reports," "The Superintendent and Testing - Implications for the Curriculum," and "Bias".

Although these and other test service publications provide for communication among educators, we are challenged to get relevant information to classroom teachers, students, parents and others in the community. It is encouraging that the new NIE Office of Testing, Assessment and Evaluation has as one of its goals the development and dissemination of easily understood materials for parents and teachers and is planning to sponsor regional workshops and conferences on critical issues in testing.

In describing a paradox in educational measurement today as one in which demands for achievement testing are increasing while the tests at the same time are being sharply criticized and opposed, Ebel (1976) discusses some of the more frequently mentioned criticisms. He concludes, "The criticism of tests is not likely to disappear, nor should it. The tests we currently use are not as good, on the whole, as they need to be. Informed criticism can help to make them better."

Test bias is an area in which criticism has contributed to the improvement of achievement tests, at least as far as face validity is concerned. Title IX legislation and the research of Tittle and others (1974) alerted authors and publishers to possible existence of sex bias and sex stereotyping in achievement and other test materials. Publishers have responded to these concerns and to charges that instructional materials, especially illustrations and story content, were not representative of the various racial, ethnic and socioeconomic groups with whom they were being used. Guidelines for identifying and eliminating sources of bias soon evolved.

In test development, careful attention is now given to the bias issue, to the writing and editing of the items and to the ways that the items function with various groups during the item tryout phases. In recent achievement test development projects, panels of minority educators have been involved and statistical techniques applied at appropriate stages. The combination of subjective and objective approaches seems to be effective in identifying and minimizing the effects of features that had given rise to earlier bias concerns.

From the focus on testing throughout the nation today, one might assume that we are facing an entirely new set of issues. To get a perspective in preparing this paper, I went back to the 1965 Review of Educational Research (Bligh, 1965) where in my review of achievement testing over a three year period I had noted trends in the measurement of educational achievement. I would now like to take

a look at several of those trends and relate them to today's concerns and to new developments in achievement testing. However, within the scope of this paper and the time limits of the symposium, it will not be possible to cover the wide range of topics that are of interest. I have selected several that seem to be especially relevant.

First, let me note that conditions since the early sixties have been favorable to the development of research in testing. The two new research journals that appeared in the early sixties, The Journal of Educational Measurement and the American Educational Research Journal are flourishing today and, with the "Validity Studies" section in Educational and Psychological Measurement, are excellent sources of information on measurement.

In the mid-sixties, objective tests were being severely criticized by Hoffman (1962) because of the multiple-choice format and heavy emphases on factual recall. Other researchers were attempting to describe the cognitive processes that achievement tests were purporting to measure. Studies were designed to demonstrate that multiple-choice items could validly measure the cognitive skills as presented in the hierarchical structure of Bloom's "Taxonomy of Educational Objectives" (1956). It may be disconcerting to hear that the multiple-choice format and the levels of skills and abilities that can be measured by standardized tests are still being seen as shortcomings. However, this is more easily understood when we consider that test users are seeking more than survey-type information. Concerns about the free-writing skills and the monitoring of instruction through pupils' open-ended responses to individual and clusters of test items present a challenge to the popular machine-scoring and processing of answer documents. For instructional purposes, where there may be little need for group summary data, hand-scorable sheets may be more practical. "Self-scoring" documents and those with latent image printing provide for more immediate feedback to the examinee.

Strong concerns that achievement tests determine the curriculum were being expressed during the mid-sixties. In an interview reported in a 1965 issue of the Grade Teacher, Roger Lennon (1965), speaking as a test publisher, voiced the concern of test makers that they not be cast in the role of shaping the curriculum. At the same time, he recognized that there was some validity in the charge that standardized tests tend to shape the school curriculum. Rudman (1977), speaking as a test author, addressed the same issues in the 1977 Phi Delta Kappa article "The Standardized Test Flap". He acknowledged that the selection of content for a major achievement series that will be meeting curriculum needs for a decade or more in the future is a "difficult and arduous" task. Through intensive analyses of new and about to be published instructional materials, achievement test authors attempt to mirror rather than to influence the curriculum.

The extent to which the achievement tests do influence or set the curriculum is difficult to assess. Undoubtedly, there is an interplay of forces. However, this concern applies not only to norm-referenced tests but perhaps even more so to tests that are designed to measure specific objectives of instruction and prescribe instruction on the basis of the assessment. A related concern is the impact that emerging instructional goals may have on existing tests and tests that are in the process of being developed. In the sixties, achievement test makers wrestled with the problems of the "new" mathematics, BSSC biology and courses of study being developed by the Physical Science Study Committee and similar groups. A somewhat anologous situation is present today as instruction in the metric system is being gradually incorporated in courses of study across the country. Selecting relevant content and identifying appropriate groups for norming tests are problems that must be resolved before the test can be published.

Standardized achievement tests are being rejected by some as not adequately meeting the needs of the local curriculum. The move toward involving teachers and

other local school personnel in the writing of objectives and constructing items to assess them sometimes results in tables of specifications that no longer can be met by published instruments. The development of special tests to meet these highly specific objectives can become costly and time consuming. Item banking and the custom building of tests is seen by some as an appropriate way to meet these demands. However, as those who have attempted to establish one well know, the development and maintenance of an item bank calls for technical knowledge and resources usually beyond the capabilities of even the larger and more affluent school systems.

The more conventional approach to achievement test development as described in the Rudman article proceeds from the assumption that in the basic skill areas there is a "universal" or "common" curriculum across the country that permits domain and content definition and the development of instructionally valid items. On several occasions, we have been asked as have others to match the objectives of our standard achievement batteries to those at a local or state level. When this is done, there is a usually high agreement among broad objectives dealing with the important cognitive skills. Further evidence of this agreement is provided by a study conducted during the fall standardization of the Metropolitan Survey and Instructional Achievement Tests. Participating teachers were asked to rate the content of the tests according to how well the tests matched their instructional objectives. Responses were then summarized by test area and level for grades K-12. At the Elementary level administered in Grade 4, for example, over 75% of the teachers surveyed indicated that they taught 116 or 94% of the 124 objectives for the reading, mathematics and language tests at that level; 90% or more of these teachers provided instruction in more than 50% of the total objectives. No single objective was taught by less than 58% of these fourth grade teachers.

In light of tightening school budgets and growing public concerns over the kinds of education being provided by the schools, I anticipate that the demand for custom-made achievement tests at the local and school level will wane or at least level off during the next decade. However, as test makers, we are challenged to provide greater flexibility in our standard test offerings.

As I have noted earlier, achievement tests today are being called upon to provide information to meet two interrelated curriculum needs -- the needs of teachers and others who want information for instructional planning and remediation, and those of administrators interested in curriculum planning and supervision, for Title I type reporting, and for communicating with the public. In addressing these needs, test makers have proceeded in several ways -- through the development of separate tests to meet individual needs, through broadening the scope of the conventional survey instrument to provide objective-referenced and item-cluster as well as survey information, and through the development of assessment systems that relate the content and psychometric properties of instructional and survey test components. The movement in recent years has been toward the latter two approaches.

Whichever of the above approaches is followed, the building of a series of achievement tests and providing information for their interpretation and use no longer ends with the date of publication. Since the mid-sixties when the survey test was often accompanied by a single manual containing information for administering, scoring and using the results, the trend has been toward providing more indepth interpretative and prescriptive information. How to do this efficiently and economically continues to be a challenge to authors and publishers who must produce and disseminate the materials and to the user who is confronted with a wealth of information from which to select. The manual for one of our recent achievement series, for example, was published in five parts including directions

for administering, norms booklets, a teacher's guide, and administrator's guide, and a report of technical data. In addition, a series of research reports, indexes of objectives and a guide for using item-analysis reports provide interpretative and technical data supplementing those given in the manuals.

In addition to the above content and curriculum concerns, we should take a brief look at some of the concerns that are more statistical in nature. Reference was made earlier to misuses of tests arising from the micrometer fallacy or the investing of test scores with a precision they were not meant to possess. During the mid-sixties, concerns about measuring growth and in establishing comparability among test scores were beginning to surface but there were relatively few studies addressing these issues. However, by 1970, the concept of accountability had emerged and there were increasing attempts within schools and by outside evaluators to administer pre- and post-tests and to assess and account for the amount of change that had taken place during the instructional period. Achievement tests were called upon to provide evidence of the changes that may have occurred. Problems soon became apparent. Wrightstone, Hogan and Abbott (Test Service Notebook No. 33) identified major issues associated with accountability in education, alerted test users to measurement problems and suggested ways of resolving them.

Responses to demands for greater precision and flexibility in achievement tests are reflected in the development of multi-level scaled-score systems and the provision of empirically derived normative information for at least two times a year, fall and spring. Adding features like this has impacted on test development in several ways -- in additional time and resources to collect the normative and scaling data, the greater involvement of school and pupils in the data-collecting procedures and added costs for generating and portraying the information.

Practical and theoretical problems in providing comparability among scores from different tests were being addressed during the mid-sixties in symposiums like this. Among the limitations cited then and still relevant today are marked differences in content and in approaches to measuring the same ability, the non-equivalence of the norm groups and the amount of error that can be safely tolerated in the interpretation of the translated scores (Angoff, 1964). There have been two important attempts, both at the national level, to provide ways of making achievement test scores comparable. These are the National Test Equating Study of Reading (The Anchor Test Study) (Jaeger, 1973) and the more recent introduction of Normal Curve Equivalents (NCEs) as alternatives to grade equivalents as a mode for Title I reporting (Tallmadge and Wood, 1976).

Latent trait models are being applied in the development of achievement tests and item data banks. Advocates of the Rasch Model see as one of its advantages, the capability of providing comparability among tests within the same content area through the calibrations of items on a common scale.

Difficulties inherent in the interpretations of test scores of groups within a school and for schools within a system were recognized in the mid-sixties. Since that time there have been several approaches to resolving these difficulties. A seemingly straightforward approach is the development of norms for groups in a fashion similar to those used in developing pupil norms. However, problems arise in defining the unit (classes, school building or system) and in drawing meaningful samples for establishing the interpretations.

A second approach has been through the provision of expectancy tables expressing relationships between achievement and ability test scores for groups as well as individual pupils. A third approach that seems to offer promise for those who may not wish to administer an ability test is through the use of socioeconomic information to interpret group performance. Based on a study of the relationships among

demographic variables reported in doctoral dissertations and other research reports and on preliminary evidence collected during item analysis, a predictive technique is being developed for the Metropolitan Achievement Tests that will relate SES data descriptive of a school building or system to the achievement levels of pupils within the building or system. The resulting reports will provide information for interpreting a group's performance in terms of the performance of groups having similar demographic characteristics.

In addition to the concerns and trends noted here, there are others that have impacted on achievement testing in various ways. Among those are requests for large city and Title I and other differentiated norms; test anxiety and the use of practice materials; the impact of scoring machines and computers on test development and the use of tests at the local level; and more direct involvement of parents in the testing of their children.

In closing, I want to call attention to an emerging demand that, if it hasn't already, is likely to become a concern to those responsible for school testing programs as well as for publishers of standardized tests. That is, the practice of open-testing or making available to students and parents copies of the test and the student's responses to the test items. In an article written for the (London) Times Educational Supplement and reprinted in the Winter issue of the edc News (Educational Development Center News), Edwin Taylor (1978) cites examples of the consequences of withholding testing information and argues for the elimination of the secrecies surrounding standardized tests.

Haney and Madaus (1978) in discussing minimal competency testing anticipate that the issue of freedom of information concerning standardized testing is likely to extend beyond the bounds of minimal competency testing. They cite the dilemma that Massachusetts' school officials now face in dealing with a State Board of Education directive that would make available to students or parents corrected tests or copies of them.

The full impact of this new demand on standardized tests remains to be felt. However, from it and other concerns still to be resolved, it appears that the next decade and a half will be as challenging as the past has been for those involved in achievement testing.

REFERENCES

ANGOFF, William H. Technical problems of obtaining equivalent scores on tests. Journal of Educational Measurement, 1964, 1, 11-13.

BLIGH, Harold F. Trends in the measurement of educational achievement. Review of Educational Research, 1965, 35, 34-52.

BLOOM, Benjamin S. (editor) Taxonomy of educational objectives: the classification of goals. Handbook 1: Cognitive Domain. New York: Longmans, Green & Co., 1956.

EBEL, Robert L. The paradox of educational testing. NCME Measurement in Education, 1976, 7, No. 4.

GLASS, Gene V. Standards and criteria. Journal of Educational Measurement, 1978, 15, 237-261.

HANEY, W. & MADAUS, G. Making sense of the competency testing movement. Harvard Educational Review, 1978, 48, 462-484.

HOFFMAN, Banesh. Toward less emphasis on multiple-choice tests. Teacher's College Record 64: 1962, 64, 183-89.

JAEGER, Richard M. The national test-equating study in reading (The Anchor Test Study). NCME Measurement in Education, 1973, 4, No. 4.

LENNON, Roger T. A test editor speaks out. Grade Teacher, April 1968.

QUINTO, F. & McKENNA, B. Alternatives to standardized testing: a handbook. Washington, D. C. National Education Association, 1977.

RUDMAN, Herbert C. The standardized test flap. Phi Delta Kappa, 1977, 59, 179-185.

TALLMADGE, G. K. & WOOD, C. T. ESEA Title I Evaluation and reporting system. User's Guide (revised), Washington, D. C., U. S. Office of Education, 1976.

TAYLOR, Edwin F. The snares of secrecy. edc News, 1978, 12.

TITTLE, C. K., McCARTHY, K., & STECKLER, J. F. Women and educational testing. Princeton, N. J., Educational Testing Service, 1974.

TURNBULL, William W. Test scores in perspective. Paper presented at the National Conference on Achievement Testing and Basic Skills, sponsored by the Department of Health, Education and Welfare, Washington, D. C., 1978.

WRIGHTSTONE, J. W., HOGAN, T. P. & ABBOTT, M. M. Accountability in education and associated measurement problems. Test Service Notebook No. 33, New York, issued by The Psychological Corporation.